Supplementary methods

Schmidt et al: Age- and tumor subtype-specific breast cancer risk estimates for *CHEK21100delC carriers**

Breast Cancer Association Consortium (BCAC) database

Data were retrieved from the BCAC database. Data are submitted by each study following the BCAC data dictionary. Central data checking, cleaning and harmonization, in communication with the study data managers and principal investigators, is done at three centers. For the core data such as case-control status and age, and for genotype data, this is performed by the group of Prof Easton (University of Cambridge), for risk factor data this is performed by the group of Prof Chang-Claude (Deutsches Krebsforschungszentrum), and for the survival, pathology and treatment data this is performed by the group of Dr Schmidt (Netherlands Cancer Institute); additional tissue microarrays data were curated by the groups of Prof Garcia-Closas (ICR) and Prof Pharoah (Cambridge University). ER, PR and HER2 status was obtained mostly from medical records followed by immunohistochemistry performed on tumor tissue microarrays or whole-section tumor slides (1).

Data freezes of the BCAC database are made for batches of data distributed for analyses. All data were used as derived from the BCAC database of August 2014 with a few exceptions. The CHEK2 genotyped data set of the ABCFS and kConFab studies included HER2 scores of one breast cancer patient each: these were recoded to missing. Missing age of 189 breast cancer patients of the ABCS study was updated. For the use of information about family history, information from the variables 'first degree female family members with breast cancer' was used to supplement the variable 'family history of breast cancer in a first degree relative (0=no, 1=yes)' for 6,149 women without and 67 women with a 63 woman first-degree female family members with breast cancer. Inclusions and exclusions of data are noted in **Supplementary figure 1** and **Supplementary tables 1-5**.

All data in the BCAC database are indexed through randomly generated unique identifiers for each individual; there are no names, addresses or places of birth in the database. The source databases at the participating institutes mostly contain coded data (anonymous for the researcher using the database), which can be indirectly or directly linked to the individuals concerned depending on the individual study protocols and approvals.

***CHEK2**1100delC genotyping**

Genotyping of *CHEK2**1100delC was done using a 5'exonuclease Taqman® allelic discrimination assay developed by the Netherlands Cancer Institute-Antoni van Leeuwenhoek hospital (NKI-AVL) using Vector NTI AlignX Invitrogen (United Kingdom) and Primer Expres 3.0 (Applied biosystems (Warrington, Cheshire, United Kingdom). Primers were specifically designed to be non-binding to the pseudogenes on chromosomes 15 and 16, which are homologous to exons 10–14 of CHEK2 on chromosome 22.

Primer and probe details

The primers developed were: Forward primer 5'-GGCAGACTATGTTAATCTTTTTATTTTATGG-3' and Reverse primer 5'-CAAGAACTTCAGGCGCCAAGT-3' (Invitrogen Ltd Paisley United Kingdom). Allele specific Minor grove binding (MGB)Probes with VIC reporter dye for Wild type allele 5'-VIC-TTTAGATTA**C**TGATTTTGGGC-MGB-3' and FAM reporter dye for the *CHEK2**1100delC allele 5'-FAM-TTAGATTATGATTTTGGGCAC-MGB-3', and Taqman Genotyping Mastermix was obtained from Applied

Biosystems (Warrington, Cheshire, United Kingdom) for the genotyping of the Dutch dataset used in the design and validation of the Taqman assay.

Validation of Taqman assay

Patients who were counseled and tested negative for *BRCA1/2* pathogenic mutations in the NKI-AVL Clinical Genetic Centre were included. In total 3,691 samples from non-*BRCA1/2* (mostly breast) cancer patients, including 347 patients with Unclassified *BRCA1/2* Variants class B1 and B2, were analyzed for *CHEK2**1100delC in the validation stage. Of this series 1,034 breast cancer cases were included in the ABCS-F (ABCS familial) sub-study in the BCAC database.

PCR reactions were run in 96 well plates on the ABI prism 7500 Fast Real-Time PCR system as follows: a hot start at 95°C for 10 min, 40 cycles; denaturation at 92°C for 15 sec and extension at 60°C for 60 sec. Each plate contained two randomly chosen duplicates, a positive and two no template controls. The call rate of the assay was 99.8% overall. All *CHEK2**1100delC mutated samples, i.e., 161 heterozygous (4.4%) and 5 homozygous (0.1%), and 65 wild type samples were validated with the P190 CHEK2 MLPA (Multiplex Ligation-dependent Probe Amplification) kit (MRC-Holland, Amsterdam, The Netherlands). There was100% concordance between the Taqman assay and the MLPA results.

An independent subset of 188 samples from the Erasmus MC in Rotterdam, which had been analyzed for *CHEK2**1100delC previously with a oligohybridisation assay (2) also showed 100% concordance with the results from the custom Taqman for 185 wild type and 3 heterozygous CHEK2*1100delC mutations.

Genotyping of BCAC samples

*CHEK2**1100delC genotyping results of 25,571 cases (with follow-up data) and 30,056 controls from 22 studies was published previously (3). As described previously (3), for the BCAC samples, a positive, negative and no template controls were included in each 96-well plate run.

Of all samples genotyped in BCAC for *CHEK2**1100delC with the custom Taqman, 3,184 called duplicates showed an overall concordance of 99% and a concordance of 92% for *CHEK2**1100delC carriers (99 of 108 carriers detected in both duplicate samples). In addition, 22,006 samples previously genotyped with other techniques (older Taqman designs, oligohybridization, and iPLEX) had been repeated with the custom designed Taqman; 33 samples were found to be discordant and were removed from the dataset, i.e. 24 false negatives (missed by older assays), 6 false positives (not confirmed *CHEK2**1100delC with custom Taqman assay), and 3 heterozygous carriers re-classified as homozygous carriers. The overall concordance rate was between the custom Taqman and the other assay was 99.8%, and the sensitivity and positive predictive value of being carrier of the 1100delC variant(s) of other assays compared to the custom Taqman were 98% and 92% respectively. Samples genotyped only with assays other than the custom Taqman were also included in the analyses. In summary, of 91,147 samples included, 80,941 samples were genotyped with the custom designed Taqman assay, 6,833 samples were only genotyped with another assay, and 3,373 were genotyped by two methods (in the latter case, the genotypes from the Taqman assay were used in preference) (**Supplementary table 1**).

Statistical analyses

*CHEK2*1100delC breast cancer risk estimates by age*

We modeled the *CHEK2**1100delC breast cancer risk estimates by age using the more stable interaction estimates for age and *CHEK2**1100delC from the case-only analysis. This analysis relies on the

assumption that *CHEK2**1100delC frequency is unrelated to age in the general population. We considered this justified since analysis in the controls showed no evidence of an association between age and *CHEK2**1100delC, and because there is little evidence for strong associations between *CHEK2**1100delC and other cancers or other phenotypes that would lead to substantial changes in frequency with age. Using a logistic regression case-control analysis, we fitted the *CHEK2* and age effects with a fixed interaction term by using the offset: chek2 x (age - 30) x ln('fixed interaction age**CHEK2* estimate from case-only analysis'). In this model, the term age-30 was used so that the main effect for *CHEK2**1100delC would correspond to the OR at age 30. Analyses were performed separately for all, ER-positive and ER-negative cases, using the fixed estimates from each respective case-only analysis.

Cumulative breast cancer risks

Cumulative risks were calculated based on estimated relative breast cancer risk for *CHEK2**1100delC carriers, using the United Kingdom breast cancer incidences 1992–2010 and the ratio ER-positive and ER-negative breast tumors from the BCAC database. Smoothed calendar period- and cohort-specific incidences were used as described previously (4). Smoothed age-specific proportions were derived from 5-year age intervals using locally weighted regression (LOWESS) with a bandwidth of 0.2. Overall and ER-specific cumulative risks in carriers were then derived from the age-specific relative risks in Figure 1 of this manuscript. The cumulative risks were not adjusted for competing risk of death before breast cancer occurrence. The cumulative risk (F_c) at age t was calculated by:

$$F_c(t) = 1 - exp\left(-\sum_{u=0}^{t-1} \lambda_0\,(u) e^{\beta(u)}\right)$$

where $\lambda_0(u)$ is the incidence rate in non-carriers at age u and $\beta(u)$ is the relative risk in *CHEK2**1100delC carriers at age u, relative to non-carriers.

$\lambda_0(u)$ was calculated such that the combined incidence in carriers and non-carriers agreed with the UK population incidences, using the formula:

$$i(u) = \lambda_0(u)\frac{QS_C(u)e^{\beta(u)}+(1-Q)S_N(u)}{QS_C(u)+(1-Q)S_N(u)}$$

where $i(u)$ is the population incidence at age u, Q is the population frequency of *CHEK2**1100delC carriers (assumed to be 0.0054), $S_N(u) = exp(-\sum_{v=0}^{u-1}\lambda_0\,(v))$ is the survival function (i.e., the probability of being unaffected) at age u in non-carriers and $S_C(u) = exp\left(-\sum_{v=0}^{u-1}\lambda_0\,(v)e^{\beta(v)}\right)$ is the corresponding survival function in carriers. $\lambda_0(u)$ $(u = 0, \ldots, 79)$ was computed iteratively using the above formulae.

Frequency rates by country

Carrier frequency estimates by country were derived using a modification of the empirical Bayes approach proposed by Clayton and Kaldor (5) for mapping disease incidence rates. This approach assumes an underlying multivariate normal distribution for the log incidence rates, and hence derives posterior estimates for the rates that account for the uncertainty in the individual estimates due to small sample size. In this application we modified the method to allow for proportions rather than rates, by assuming that logit of the proportion had an underlying normal distribution. To utilize the data from cases and controls, we first obtained bias-corrected log(odds) estimates separately for cases and controls in each country, and combined these using an invariance variance weighting, and offsetting the case estimate by the log-odds ratio for *CHEK2* status in the analysis to allow for the higher frequency in

cases. This estimate was then used as the basis for deriving the approximate likelihood given by Clayton and Kaldor, and hence deriving the empirical estimates by an expectation–maximization algorithm. Correlations in the frequencies among countries were allowed for by assuming a conditional autoregression model (6) with a correlation ρ between neighboring countries. In this analysis, however, the estimate for ρ converged to zero.

References:

1) Broeks A#, Schmidt MK#, Sherman ME#, Couch FJ, Hopper JL, Dite GS, Apicella C, Smith LD, Hammet F, Southey MC, Van 't Veer LJ, de Groot R, Smit VT, Fasching PA, Beckmann MW, Jud S, Ekici AB, Hartmann A, Hein A, Schulz-Wendtland R, Burwinkel B, Marme F, Schneeweiss A, Sinn HP, Sohn C, Tchatchou S, Bojesen SE, Nordestgaard BG, Flyger H, Orsted DD, Kaur-Knudsen D, Milne RL, Pérez JI, Zamora P, Rodríguez PM, Benítez J, Brauch H, Justenhoven C, Ko YD; The Genica Network, Hamann U, Fischer HP, Brüning T, Pesch B, Chang-Claude J, Wang-Gohrke S, Bremer M, Karstens JH, Hillemanns P, Dörk T, Nevanlinna HA, Heikkinen T, Heikkilä P, Blomqvist C, Aittomäki K, Aaltonen K, Lindblom A, Margolin S, Mannermaa A, Kosma VM, Kauppinen JM, Kataja V, Auvinen P, Eskelinen M, Soini Y, Chenevix-Trench G, Spurdle AB, Beesley J, Chen X, Holland H; kConFab; AOCS, Lambrechts D, Claes B, Vandorpe T, Neven P, Wildiers H, Flesch-Janys D, Hein R, Löning T, Kosel M, Fredericksen ZS, Wang X, Giles GG, Baglietto L, Severi G, McLean C, Haiman CA, Henderson BE, Le Marchand L, Kolonel LN, Grenaker Alnæs G, Kristensen V, Børresen-Dale AL, Hunter DJ, Hankinson SE, Andrulis IL, Marie Mulligan A, O'Malley FP, Devilee P, Huijts PE, Tollenaar RA, Van Asperen CJ, Seynaeve CS, Chanock SJ, Lissowska J, Brinton L, Peplonska B, Figueroa J, Yang XR, Hooning MJ, Hollestelle A, Oldenburg RA, Jager A, Kriege M, Ozturk B, van Leenders GJ, Hall P, Czene K, Humphreys K, Liu J, Cox A, Connley D, Cramp HE, Cross SS, Balasubramanian SP, Reed MW, Dunning AM, Easton DF, Humphreys MK, Caldas C, Blows F, Driver K, Provenzano E, Lubinski J, Jakubowska A, Huzarski T, Byrski T, Cybulski C, Gorski B, Gronwald J, Brennan P, Sangrajrang S, Gaborieau V, Shen CY, Hsiung CN, Yu JC, Chen ST, Hsu GC, Hou MF, Huang CS, Anton-Culver H, Ziogas A, Pharoah PD, Garcia-Closas M#. Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the Breast Cancer Association Consortium. Hum Mol Genet. 2011 Aug 15;20(16):3289-3303.

2) Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, Hollestelle A, Houben M, Crepin E, van Veghel-Plandsoen M, Elstrodt F, van Duijn C, Bartels C, Meijers C, Schutte M, McGuffog L, Thompson D, Easton D, Sodha N, Seal S, Barfoot R, Mangion J, Chang-Claude J, Eccles D, Eeles R, Evans DG, Houlston R, Murday V, Narod S, Peretz T, Peto J, Phelan C, Zhang HX, Szabo C, Devilee P, Goldgar D, Futreal PA, Nathanson KL, Weber B, Rahman N, Stratton MR; CHEK2-Breast Cancer Consortium. Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. Nat Genet. 2002 May;31(1):55-9.

3) Weischer M, Nordestgaard BG, Pharoah P, Bolla MK, Nevanlinna H, Van't Veer LJ, Garcia-Closas M, Hopper JL, Hall P, Andrulis IL, Devilee P, Fasching PA, Anton-Culver H, Lambrechts D, Hooning M, Cox A, Giles GG, Burwinkel B, Lindblom A, Couch FJ, Mannermaa A, Grenaker Alnæs G, John EM, Dörk T, Flyger H, Dunning AM, Wang Q, Muranen TA, van Hien R, Figueroa J, Southey MC, Czene K, Knight JA, Tollenaar RA, Beckmann MW, Ziogas A, Christiaens MR, Collée JM, Reed MW, Severi G, Marme F, Margolin S, Olson JE, Kosma VM, Kristensen VN, Miron A, Bogdanova N, Shah M, Blomqvist C, Broeks A, Sherman M, Phillips KA, Li J, Liu J, Glendon G, Seynaeve C, Ekici AB, Leunen K, Kriege M, Cross SS, Baglietto L, Sohn C, Wang X, Kataja V, Børresen-Dale AL, Meyer A, Easton DF, Schmidt MK, Bojesen SE. CHEK2*1100delC

Heterozygosity in Women With Breast Cancer Associated With Early Death, Breast Cancer-Specific Death, and Increased Risk of a Second Breast Cancer. J Clin Oncol. 2012 Dec 10;30(35):4308-16.

4) Lee AJ, Cunningham AP, Kuchenbaecker KB, Mavaddat N, Easton DF, Antoniou AC, Consortium of Investigators of Modifiers of B, Breast Cancer Association C: BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. Br J Cancer 2014, 110(2):535-545.

5) Clayton D, Kaldor J: Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 1987, 43(3):671-681.

6) Besag JE: Spatial interaction and the statistical analysis of lattice systems. . Journal of the Royal Statistical Society 1974, Series B36:192-236.